Interkulturell vergleichende Umfragen

Braun, Michael

Interkulturell vergleichende Umfragen

56

Michael Braun

Die interkulturell vergleichende Umfrageforschung hat in den Sozialwissenschaften in den letzten Jahrzehnten stark an Bedeutung gewonnen. Dies betrifft zum einen die Verfügbarkeit von international vergleichbaren Umfragedaten und die auf ihnen basierenden substanzwissenschaftlichen Untersuchungen, zum anderen die Beschäftigung mit methodologischen Fragestellungen. Für interkulturelle Umfragen sind nicht nur die Probleme, die für alle Umfragen allgemein gelten, zu berücksichtigen. Vielmehr bemisst sich die Qualität eines interkulturellen Projekts an der Qualität der einzelnen nationalen Umfragen und insbesondere auch daran, dass diese „vergleichbar" sind. Dabei spielt es zunächst keine Rolle, ob unterschiedliche ethnische Gruppen in einem einzigen Land oder unterschiedliche Länder miteinander verglichen werden sollen. Bei der Untersuchung verschiedener ethnischer Gruppen (El-Menouar, Kapitel 60 in diesem Band) innerhalb eines Landes sind in der Regel z.B. die Stichprobenziehung (Häder/Häder, Kapitel 18 in diesem Band) und die Übersetzungsproblematik zu berücksichtigen. Bei einem Vergleich unterschiedlicher Länder kommen noch unterschiedliche institutionelle Randbedingungen im weitesten Sinn hinzu (von den Bedingungen der Stichprobenziehung zu den sozioökonomischen und politischen Gegebenheiten). Wegen der größeren Allgemeinheit erfolgt im Folgenden eine Beschränkung auf den Vergleich unterschiedlicher Länder, der Vergleich verschiedener kultureller Gruppen ist aber in der Regel ebenso angesprochen. Zunächst stellt sich die Frage, welche Länder überhaupt in ein Umfrageprojekt oder die Analyse einbezogen werden sollen.

56.1 Auswahl der untersuchten Länder oder kultureller Gruppen

Die Auswahl der zu vergleichenden Länder oder kulturellen Gruppen spielt bei dem Design eines Forschungsprojektes und der Analyse von Sekundärdaten (Mochmann, Schupp, Kapitel 14 und 73 in diesem Band) eine Rolle. Hierfür bieten sich je nach Forschungsfrage

unterschiedliche Strategien an (z.B. Przeworski/Teune 1970, Scheuch 1968): Unähnliche *Kontextbedingungen*, d.h. die Auswahl möglichst unterschiedlicher Länder, sind ein Vorzug, wenn die Generalisierbarkeit von Zusammenhängen zwischen den untersuchten Variablen aufgezeigt werden soll, wie im „most different systems"-Design (Przeworski/Teune 1970: 34ff). Stellt sich dann heraus, dass diese Zusammenhänge in den untersuchten Ländern tatsächlich gleich sind, besteht begründeter Anlass, von deren genereller Geltung auszugehen. Es kann dann nämlich weitgehend ausgeschlossen werden, dass die in einem Land gefundenen Zusammenhänge nur bei einem Zusammentreffen ganz spezifischer Kontextbedingungen bestehen. Die *Systemebene*, d.h. die Merkmale der einzelnen Länder, wird nur dann als Erklärungsfaktor berücksichtigt, wenn die Daten der Annahme der Generalisierbarkeit widersprechen, d.h. wenn in den einzelnen Ländern unterschiedliche Zusammenhänge gefunden werden. Soll demgegenüber nachgewiesen werden, dass eine in einem Land oder einer kulturellen Gruppe gefundene Beziehung nicht überall gilt, erweist sich die Verwendung unterschiedlicher Kontexte (d.h. von Ländern, die sich in einer Vielzahl von Merkmalen unterscheiden) als Nachteil. Wenn die Daten nämlich der Erwartung unterschiedlicher Zusammenhänge scheinbar entsprechen, können Alternativerklärungen eine eindeutige Interpretation unmöglich machen. In diesem Fall ist es vielmehr sinnvoll, möglichst ähnliche Länder auszuwählen (das „most similar systems"-Design von Przeworski/Teune 1970). Werden nämlich schon bei Ländern, die sich hinsichtlich vieler relevanter Merkmale ähnlich sind, Unterschiede in den Zusammenhängen gefunden, so ist davon auszugehen, dass dies erst recht bei sehr verschiedenartigen Ländern der Fall sein wird.

In der Forschungspraxis verlieren diese Überlegungen allerdings zunehmend an Relevanz, zumindest was die Auswahl von Ländern anbetrifft. Viele Projekte der international vergleichenden Umfrageforschung, wie z.B. das International Social Survey Program (ISSP, http://www.issp.org/) oder der World Values Survey (WVS, http://www.worldvaluessurvey.org/) bemühen sich um eine immer umfassendere Abdeckung aller Länder. Auch die auf einzelne Weltregionen beschränkten Umfrageprojekte wie der European Social Survey (ESS, http://www.europeansocialsurvey.org/), die European Values Study (EVS, http://www.europeanvaluesstudy.eu/) oder die Eurobarometer-Umfragen (http://ec.europa.eu/public_opinion/index_en.htm) verfolgen in ihrem Bereich eine möglichst komplette Abdeckung der existierenden Länder (zu diesen Datensätzen siehe Mochmann, Kapitel 14 in diesem Band). Zudem existiert mit der Mehrebenenanalyse (Pötschke, Kapitel 87 in diesem Band) ein für interkulturelle Vergleiche besonders geeignetes statistisches Verfahren, mit welchem eine große Zahl von Ländern verglichen werden kann. Mit diesem Verfahren kann z.B. untersucht werden, ob Unterschiede in den Zusammenhängen zwischen den einzelnen Ländern bestehen und wie diese gegebenenfalls mit Merkmalen auf der Länderebene erklärt werden können.

56.2 Probleme der Vergleichbarkeit der Projektkomponenten

In der interkulturell vergleichenden Forschung hängen die Schlussfolgerungen von der Qualität und Vergleichbarkeit der einzelnen nationalen Studien ab. Sind einige von ihnen mit Fehlern behaftet, dann entsprechen beobachtete Ähnlichkeiten und Unterschiede möglicherweise lediglich methodischen Artefakten. Dabei kann auch die Verwendung identischer Prozeduren in den einzelnen Ländern keine Vergleichbarkeit garantieren. Teilweise ist die Verwendung unterschiedlicher, aber den jeweiligen Kontexten angepasster Untersuchungsstrategien vorzuziehen. So macht es offensichtlich wenig Sinn, die Durchführung von Telefonumfragen (Hüfken, Kapitel 46 in diesem Band) für alle Länder verbindlich vorzuschreiben. In Ländern mit sehr geringer Telefondichte würde dadurch ein großer Teil der Bevölkerung von vorneherein von der Umfrage ausgeschlossen. Stattdessen sollte man in diesen Ländern auf einen anderen Erhebungsmodus ausweichen, also auf eine persönlich-mündliche (Stocké, Kapitel 45 in diesem Band) oder schriftliche Befragung (Reuband, Kapitel 47 in diesem Band). Auch bei der Stichprobenziehung wird man Unterschiede in der konkreten Durchführung akzeptieren müssen, wenn man die in den einzelnen Ländern jeweils besten Verfahren anwenden möchte. So gilt in Deutschland etwa eine aus den Registern der Einwohnermeldeämter gezogene Stichprobe (Häder/Häder, Kapitel 18 in diesem Band) als optimal für persönlich-mündliche Befragungen. In den Vereinigten Staaten gibt es solche Register allerdings nicht. Dort besteht das Standardverfahren bei wissenschaftlichen Umfragen darin, zuerst zufällig eine Stichprobe von Häuserblöcken zu ziehen, in den ausgewählten Häuserblöcken dann eine weitere Stichprobe von Haushalten und in den ausgewählten Haushalten schließlich die zu befragenden Personen mit einem Zufallsverfahren zu ermitteln. In einigen Entwicklungsländern ist auch ein solches Verfahren nicht durchführbar, z.B. weil keine belastbaren Informationen über die (kleinräumige) Verteilung der Bevölkerung vorliegt, Häuserblocks nicht identifizierbar sind oder ein Teil der Bevölkerung aus Nomaden besteht. In diesen Fällen muss auf weniger stringente Vorgehensweisen zurückgegriffen werden, die in der Regel keine Zufallsstichproben mehr gewährleisten. Nun wäre es kaum vertretbar, nur um der Gleichheit des Vorgehens Genüge zu tun, solche suboptimalen Verfahren auch in den Ländern einzusetzen, in denen gute Zufallsstichproben gezogen werden können. Schließlich müssen auch bei der Formulierung der Items länderspezifische Besonderheiten beachtet werden. So hat der Staatspräsident in den Vereinigten Staaten eine andere Funktion als in Deutschland. Sollen etwa Einstellungen zu dem jeweiligen Regierungsoberhaupt erfasst werden, müsste daher in der deutschen Fragebogen-Version „President" durch „Bundeskanzlerin" ersetzt werden.

Aus diesen Beispielen wird auch deutlich, dass man zwei große Klassen von Fehlern unterscheiden kann (Faulbaum, Kapitel 31 in diesem Band): Erstens repräsentieren die in den unterschiedlichen Ländern gezogenen und realisierten Stichproben die jeweiligen Grundgesamtheiten in vergleichbarer Weise (*coverage und non-reponse error*) (Engel/Schmidt, Kapitel 23 in diesem Band)? Zweitens funktionieren die Messinstrumente, d.h. die Fragebögen und die einzelnen Fragen nach der Übersetzung, überall gleich (*Messfehler*) (Krebs/Menold, Kapitel 30 in diesem Band)?

Ein „coverage error" (Häder/Häder, Kapitel 18 in diesem Band) liegt vor, wenn nicht alle Einheiten der zu untersuchenden Population eine von Null verschiedene Wahrscheinlichkeit aufweisen, in die Stichprobe zu gelangen. Der „non-response error" (Engels/Schmidt, Kapitel 23 in diesem Band) bezieht sich auf die Nicht-Teilnahme von für die (Brutto-) Stichprobe ausgewählten Einheiten der Grundgesamtheit. Er setzt sich zusammen aus dem Anteil der Nicht-Erreichten, der Nicht-Kooperativen und der aus anderen, wie z.B. gesundheitlichen oder sprachlichen Gründen nicht befragungsfähigen Personen. Alle Komponenten können interkulturell stark variieren, z.B. in Abhängigkeit vom Umfrageklima in einer Gesellschaft (d.h. der Wertschätzung von Umfragen und der Bereitschaft, an ihnen teilzunehmen), von der Mobilität der Bevölkerung, der durchschnittlichen Haushaltsgröße, den verfügbaren personellen, finanziellen und organisatorischen Ressourcen des Umfrageinstituts oder der Art des Auftraggebers. Die Auswirkungen des Ausschlusses bestimmter Bevölkerungsteile – sei es durch coverage oder non-response error – hängen unter anderem vom Thema der Umfrage und von der Größe der betroffenen Gruppen ab sowie davon, wie stark sich diese Gruppen von dem in die Umfrage einbezogenen Bevölkerungsteil unterscheiden. Als Messfehler („measurement error") wird eine Reihe von miteinander interagierenden Fehlerquellen bezeichnet, die auf das Messinstrument (Reinecke, Kapitel 44 in diesem Band), die Interviewer (Michael/Glantz, Kapitel 21 in diesem Band), die Befragten oder den Umfragemodus zurückgeführt werden können.

56.3 Das Problem der Äquivalenz von Konstrukten und Items

Die Vergleichbarkeit von Konstrukten und Items über die unterschiedlichen Länder hinweg wird in der Regel als „Äquivalenz" bezeichnet. Dabei lassen sich mehrere Typen von Äquivalenz unterscheiden (van de Vijver/Leung 1997): Konstruktäquivalenz („construct equivalence"), Äquivalenz der Maßeinheit („measurement unit equivalence") und skalare Äquivalenz („scalar equivalence").

Interkulturell vergleichenden Studien haben unabhängig von ihren Zielen und der Art der erhobenen Daten eine unverzichtbare Voraussetzung: die *Konstruktäquivalenz* (z.B. Scheuch 1968). Sinnvoll kann in jedem Fall nur dann verglichen werden, wenn in den unterschiedlichen Ländern dieselbe zugrunde liegende Dimension erfasst wird. Ist dies nicht der Fall, haben die Konstrukte in den einzelnen Ländern eine unterschiedliche Bedeutung. Möglicherweise existieren in einigen Ländern für ein theoretisches Konstrukt oder die konkreten, bei der Messung verwendeten Begriffe keine Entsprechungen in der Realität. Dann ist aber überhaupt kein internationaler Vergleich mehr möglich. Ein häufig genanntes Beispiel ist das Konzept der „filial piety", d.h. der Erwartungen an einen guten Sohn oder eine gute Tochter. In kollektivistischen Gesellschaften (wie z.B. China) schließen diese Erwartungen in stärkerem Maße als in individualistischen Gesellschaften (wie z.B. Deutschland oder die Vereinigten Staaten) die Betreuung der eigenen Eltern im Alter ein. Ein anderes Beispiel wäre ein Vergleich von Einstellungen zu Geschlechterrollen, bei denen in der Regel Vorstellungen und Einstellungen zur Erwerbstätigkeit der Frauen eine

wichtige Rolle spielen. In weitgehend agrarisch geprägten Gesellschaften, in denen Frauen (und auch Männer) selten außerhalb des Haushaltes arbeiten, werden entsprechende Fragen möglicherweise anders (oder aber gar nicht) verstanden als in Gesellschaften, in denen Erwerbstätigkeit räumlich vom Familienleben getrennt ist.

Konstruktäquivalenz ist auch die Voraussetzung für die Äquivalenz der Maßeinheit und der skalaren Äquivalenz. Die Sicherstellung von Konstruktäquivalenz kann einen Einsatz verschiedener Messinstrumente in unterschiedlichen Kulturen erfordern (siehe dazu den nächsten Abschnitt). Dann sind allerdings die beiden anderen Äquivalenzarten möglicherweise nicht mehr gegeben.

Äquivalenz der Maßeinheit bedeutet, dass die Maßeinheit der Skala (Franzen, Kapitel 51 in diesem Band) in unterschiedlichen Kulturen gleich ist. Konkret bedeutet das, dass der Unterschied zwischen zwei Skalenpunkten, z.B. „unzufrieden" und „ziemlich unzufrieden" bei einer Frage zur Zufriedenheit mit der Arbeit in den einzelnen Ländern auch gleichen Unterschieden auf der zugrundeliegenden (nicht direkt messbaren) Dimension entspricht. Für den Ursprung der Skala ist aber eine solche Gleichheit zwischen den Ländern dann nicht noch nicht unbedingt gegeben. So könnte in einem Land die Beantwortung der Frage durch einen externen Störfaktor in eine bestimmte Richtung verzerrt werden, zum Beispiel durch eine Tendenz, extreme Antworten generell zu vermeiden. In diesem Land würde dann z.B. die Antwortalternative „völlig unzufrieden" überhaupt nicht verwendet, auch nicht von denen, die tatsächlich völlig mit ihrer Arbeit zufrieden sind (und somit den gleichen „Nullpunkt" auf der entsprechenden Dimension aufweisen wie diejenigen, die in einem anderen Land mit „völlig unzufrieden" antworten. Obwohl dies nichts mit der eigentlich zu messenden Dimension zu tun hat, beeinflusst dies die Ergebnisse. Damit sind auch die gefundenen Mittelwertunterschiede zwischen den einzelnen Ländern irreführend, da sie ja neben der eigentlich zu messenden Dimension auch noch externe Störfaktoren enthalten.

Skalare Äquivalenz bedeutet demgegenüber, dass sowohl Maßeinheit als auch Ursprung der Skala gleich sind. Letzteres setzt voraus, um bei dem Beispiel der Arbeitszufriedenheit zu bleiben, dass in allen Ländern diejenigen, die völlig mit ihrer Arbeit unzufrieden sind, bei der entsprechenden Frage auch „völlig unzufrieden" angeben (oder aber auch nur „sehr unzufrieden", solange das nur in allen Ländern gleich geschieht). Dann haben Personen in verschiedenen Ländern, die die gleiche Ausprägung auf der zugrundeliegenden Dimension haben, auch den gleichen Nullpunkt bei der für die Messung der Dimension verwendeten Frage. Erst wenn skalare Äquivalenz gegeben ist, können auch die Unterschiede in den Mittelwerten der Fragen und Items über die einzelnen Länder hinweg inhaltlich interpretiert werden.

Wie bereits erwähnt, kann ein Einsatz kulturspezifischer Messinstrumente erforderlich sein, wenn identische Fragen unabhängig von einer möglicherweise fehlerhaften Übersetzung in verschiedenen Ländern entweder überhaupt nicht oder unterschiedlich verstanden würden, und zwar nicht nur aufgrund der Formulierung, sondern auch aufgrund des Realitätsbezugs von Items. Das von Przeworski/Teune (1966) vorgeschlagene „identity-equivalence"-Verfahren ist in solchen Situationen einsetzbar: Zunächst werden die

für alle Länder identischen Indikatoren bestimmt und mit den jeweils kulturspezifischen verbunden. Daraus resultieren unterschiedliche Skalen für die einzelnen Länder. Mithilfe der identischen Items lässt sich dann die funktionale Äquivalenz der kulturspezifischen Items überprüfen. Insgesamt dürften in der Politikwissenschaft stark kontextabhängige Messungen, die sich nicht durch identische Instrumente durchführen lassen, häufiger sein als in anderen sozialwissenschaftlichen Bereichen (Przeworski/Teune 1970). Allerdings ist dies dort auch offensichtlicher, was die Problemlösung erleichtern kann: Bei hoher Kontextabhängigkeit ist von vorneherein klar, dass kein identisches Instrument konstruiert werden kann. Bei mittlerer Kontextabhängigkeit fehlt nicht nur meist die Bereitschaft der Forscher, für die einzelnen Länder manifest unterschiedliche, aber funktional äquivalente Messinstrumente zu konstruieren. Häufig ist auch nicht einfach bestimmbar, wie solche Varianten einer Frage auszusehen haben, mit denen in allen Ländern das Gleiche erfasst werden kann. Die Lösung besteht dann oft darin, sich auf den kleinsten gemeinsamen Nenner zu beschränken. Ein Messinstrument wird also so konstruiert, dass es in allen Ländern eingesetzt werden kann, wobei eine möglicherweise unvollständige Abdeckung der eigentlich intendierten Dimension in Kauf genommen wird. So lassen sich etwa beim Thema Religion nicht alle für christliche Gesellschaften sinnvollen Aspekte abfragen, wenn sie in anderen Religionen keine Entsprechung haben.

56.4 Herstellung der Äquivalenz der Messinstrumente

Notwendig ist die Herstellung einer Vergleichbarkeit der Stimuli (Hoffmeyer-Zlotnik/Warner, Kapitel 54 in diesem Band), die die Befragten letztendlich verarbeiten. Ist diese Vergleichbarkeit nicht gegeben, reflektieren die Daten nicht nur reale Unterschiede zwischen den untersuchten Kulturen, sondern auch Artefakte der Messung. Die Trennung zwischen beiden stellt eine große Herausforderung dar. Es ist natürlich sinnvoll, durch eine sorgfältige Konstruktion der Erhebungsinstrumente ex-ante sämtliche Inäquivalenzen zu vermeiden. Dies ist jedoch in vielen Studien in der Vergangenheit trotz großer Anstrengungen nicht vollständig gelungen. Daher bleibt für den Nutzer von Sekundärdaten die Aufgabe, Äquivalenz ex-post zu überprüfen.

56.4.1 Herstellung der Äquivalenz der Messinstrumente ex-ante

Die Vergleichbarkeit der durch bestimmte Messinstrumente erhobenen Daten kann sowohl durch eine unangemessene Übertragung des Fragebogens bzw. einzelner Items in andere Sprachen als auch durch Unterschiede in der sozialen Realität in den unterschiedlichen Ländern beeinträchtigt werden. Die Herstellung von funktionaler Äquivalenz ex-ante wird in der Regel durch eine sorgfältige Konstruktion des Source-Fragebogens unter Einbezug der vorhandenen interkulturellen Kompetenz sowie eine fachgerechte Übersetzung und Adaption von Fragebögen angestrebt. Beispielsweise wird beim ISSP zunächst

ein englischsprachiger Fragebogen von einer sogenannten Drafting Group vorbereitet, die aus den Vertretern von etwa einem halben Dutzend möglichst unterschiedlicher Länder besteht. In einem iterativen Verfahren wird dieser Entwurf mit den Vertretern der übrigen Mitgliedsländer des ISSP diskutiert, in einer kleineren Anzahl von Ländern getestet und schließlich durch das ISSP-Plenum nach ausführlicher Diskussion und Abstimmung verabschiedet. In diesem Prozess werden länderspezifische Besonderheiten ausdrücklich berücksichtigt, sowohl was die Einbeziehung einzelner Items als auch deren Formulierung betrifft. So haben etwa ostasiatische Länder darauf hingewirkt, dass Fragen zur Religion auf einem etwas höheren Abstraktionsniveau formuliert wurden, als es nötig gewesen wäre, wenn alle Länder einen christlichen Hintergrund gehabt hätten. Anschließend wird der Source-Fragebogen in die Sprachen der beteiligten Länder übersetzt. Empfohlen wird hierbei ein Teamansatz, bei dem zwei Übersetzer den Fragebogen zunächst unabhängig voneinander in die Landessprache übertragen und dann gemeinsam mit Experten für das betreffende inhaltliche Thema und Umfrageexperten besprechen. Dabei wird besonders berücksichtigt, ob die Items in der Übersetzung genau so verstanden werden wie im Ausgangsfragebogen. Die daraus hervorgehende landessprachliche Fragebogen-Version wird einem (kognitiven) Pretest unterzogen, wobei insbesondere die Übersetzung überprüft wird, also ob es Probleme mit der Verständlichkeit gibt und ob tatsächlich jeweils die intendierten Dimensionen gemessen werden.

Die Unterschiede zwischen Übersetzung und Adaption sind fließend. Teils wird zwischen beiden Prozessen nicht deutlich unterschieden, obwohl dies sinnvoll wäre. Übersetzung (im engeren Sinne) meint die sprachlichen Aspekte. Da aber auch sprachlich gut übersetzte Fragen in verschiedenen Kulturen unterschiedliche Stimuli darstellen können, ist zusätzlich eine Adaption, die auf (nicht-sprachliche) kulturelle Besonderheiten Rücksicht nimmt, erforderlich. Fehler werden hierbei paradoxerweise nicht nur dort vorkommen, wo die interkulturellen Unterschiede offensichtlich sind, sondern gerade auch dort, wo sie scheinbar geringer sind und leicht übersehen werden können (z.B. wenn wegen der gemeinsamen Sprache keine Übersetzung erforderlich ist). So konnte etwa gezeigt werden, dass bei Fragen zu den Auswirkungen der Berufstätigkeit einer Mutter auf ihre Kinder ostdeutsche Befragte eher an jüngere Kinder und einen größeren Umfang der Berufstätigkeit der Mutter denken als westdeutsche Befragte. Die Angaben der Befragten sind damit nicht direkt vergleichbar.

Als Hilfsmittel bei der Sicherstellung von Äquivalenz ex-ante bieten sich qualitative Verfahren und quantitative Pretests an. Bei den qualitativen Verfahren dominieren kognitive Interviews, bei denen Probleme bei der Beantwortung der Fragen herausgearbeitet werden sollen (Willis 2005). *International* vergleichende kognitive Interview-Studien sind allerdings im Vergleich zu *interkulturell* vergleichenden kognitiven Interview-Studien in einem Land wegen des hohen Koordinationsaufwands sehr selten (als eine Ausnahme: Miller et al. 2011). Bei quantitativen Pretests wird der für die Hauptstudie vorgesehene Fragebogen unter – mehr oder weniger – realen Bedingungen getestet. Wenn Pretests in den unterschiedlichen Ländern durchgeführt werden und dabei entsprechend hohe Fallzahlen realisiert werden, können auch statistische Verfahren zur Äquivalenzüberprüfung eingesetzt werden.

In der Realität tritt die Situation, dass in allen beteiligten Ländern große quantitative Pretests oder qualitative Vorstudien durchgeführt werden, wegen des hohen Aufwandes praktisch nicht auf. Selbst für den European Social Survey (ESS) waren qualitative Studien die Ausnahme und ein quantitativer Pretest wurde vor der Verabschiedung des englischsprachigen Source-Fragebogens nur in zwei der beteiligten Länder durchgeführt. Zudem werden diese Pretests – wie auch beim ISSP – vor allem unter dem Gesichtspunkt der Auswahl von inhaltlich interessanten und methodisch adäquaten Fragen durchgeführt und nicht im Sinne eines Tests des endgültigen, auch in der Hauptstudie verwendeten Messinstruments. Daher ist es oft erforderlich, zusätzlich eine Überprüfung der Äquivalenz ex-post, d.h. auf der Grundlage der in der Hauptstudie erhobenen Daten, vorzunehmen.

56.4.2 Überprüfung der Äquivalenz der Messinstrumente ex-post

Zur Überprüfung von Äquivalenz ex-post gibt es eine Vielzahl quantitativer Verfahren. Das wohl mit Abstand am häufigsten angewandte Verfahren ist die konfirmatorische Faktorenanalyse. Braun/Johnson (2011) wenden eine ganzer Reihe komplexer und weniger komplexer Verfahren auf das gleiche inhaltliche Problem an und können dadurch zeigen, dass sie im Wesentlichen zu den gleichen Schlussfolgerungen führen (siehe dort auch Literatur zu den einzelnen Verfahren).

Ein Nachteil aller quantitativen Verfahren ist aber, dass sie zwar das Vorhandensein von Problemen feststellen, deren eigentliche Ursachen aber nicht erklären können. Hier bieten sich dann – wie bereits bei der Sicherstellung von Äquivalenz ex-ante (also im Rahmen von Pretests) Probingtechniken an. Da die Durchführung kognitiver Interviews – zumal im internationalen Vergleich – äußerst aufwendig ist, erscheint hier die Durchführung zusätzlicher webbasierter Studien eine mögliche Vorgehensweise. Dabei können Probingfragen – entsprechend Schumans (1966) Vorschlag für „random probes" – in einen normalen Web-Fragebogen eingeschaltet werden.

Behr et al. (2012) nutzen beispielsweise zur Analyse eines Items zu zivilem Ungehorsam aus dem ISSP sowohl „comprehension" („Was verbinden Sie mit dem Begriff ‚ziviler Ungehorsam', nennen Sie Beispiele") als auch „category-selection probes" („Bitte begründen Sie, warum Sie sich für [Antwortalternative] entschieden haben"). Sie können dabei zeigen, dass die geringe Befürwortung von zivilem Ungehorsam in Ländern wie den Vereinigten Staaten und Kanada im Vergleich zu Ländern wie Deutschland und Spanien im Wesentlichen zwei Gründe hat. Zum einen wird in der ersten Gruppe von Ländern ziviler Ungehorsam stärker mit Gewalt assoziiert als in der zweiten. Dies ist ein Methodenartefakt, d.h. die Befragten beantworten faktisch unterschiedliche Fragen. Zum anderen ist das Vertrauen in die Politiker in der zweiten Gruppe von Ländern (noch) geringer als in der ersten. Dies ist ein inhaltliches Ergebnis. In den Daten sind aber faktisch Methodenartefakte und reale Unterschiede konfundiert, die nicht (oder nur sehr schwer) unterschieden werden können.

Braun et al. (2012) können zeigen, dass Befragte aus unterschiedlichen Ländern bei Fragen zu Migranten zwar weitgehend an vergleichbare, der tatsächlichen Migrationsrealität entsprechende Gruppen denken, dass es dabei aber dennoch zu charakteristischen Verzerrungen kommt. Sie verwenden dazu „specific probes" („An welche Gruppen haben Sie bei der Beantwortung der Frage gedacht?"). Ob die Einstellungen der Befragten verschiedener Länder zu Migranten möglicherweise deshalb nicht sinnvoll miteinander verglichen werden können, weil trotz vergleichbarer Migrationsrealität an unterschiedliche Gruppen gedacht wird, hätte mit den herkömmlichen statistischen Verfahren alleine nicht untersucht werden können.

56.5 Weitere Aspekte und weiterführende Literatur

Die zuvor aufgeführten Fehlerquellen betreffen alle in Umfragen erhobenen Variablen, z.B. demographische Informationen, Verhaltensberichte und Einstellungsfragen. Demographische Fragen können z.B. genauso falsch übersetzt werden wie Einstellungsfragen. Auch die Abfrage von objektiven Informationen kann Verständnis- und Interpretationsschwierigkeiten aufwerfen und mit Schwierigkeiten beim Abruf der gewünschten Informationen aus dem Gedächtnis verbunden sein. Möglicherweise unterschiedliche Definitionen in den verschiedenen Ländern betreffen nahezu alle sozio-demographischen Variablen (Hoffmeyer-Zlotnik/Wolf 2003, Scheuch 1968). Bei einigen dieser Variablen, z.B. bei der Bildung, werden die Probleme noch durch die große Unterschiedlichkeit der zugrundeliegenden Systeme verschärft.

Das Äquivalenzproblem betrifft im Übrigen nicht nur die interessierenden Variablen, sondern auch die Auswahl von zu vergleichenden Gruppen in den einzelnen Ländern. Bei internationalen Vergleichen ist stets zu berücksichtigen, ob vielleicht gar keine Unterschiede zwischen der gleichen Gruppe hinsichtlich einer Zielvariablen in verschiedenen Ländern festgestellt worden sind, sondern stattdessen unterschiedliche Gruppen miteinander verglichen wurden. Scheuch (1968) diskutiert beispielsweise, ob Bauern in den USA mit denen in Europa vergleichbar sind, ob man also bei einem interkulturellen Vergleich der Antworten nicht ähnliche, sondern verschiedene Gruppen miteinander vergleicht.

In den letzten Jahren sind eine ganze Reihe von Sammelbänden, Monographien und Einzelartikeln erschienen, die sich mit den methodologischen Problemen der interkulturell vergleichenden Umfrageforschung beschäftigen, z.B. Davidov et al. (2011), Harkness et al. (2010), Harkness et al. (2003), Hoffmeyer-Zlotnik/Wolf (2003), van Deth (2013[1998]) sowie van de Vijver/Leung (1997).

Literatur

Behr, Dorothée/Braun, Michael/Kaczmirek, Lars/Bandilla, Wolfgang (2012): Item comparability in cross-national surveys: Results from asking probing questions in cross-national Web surveys about attitudes towards civil disobedience. In: Quality & Quantity. 2012. DOI: 10.1007/s11135-012-9754-8

Braun, Michael/Behr, Dorothée/Kaczmirek, Lars (2012): Assessing cross-national equivalence of measures of xenophobia: Evidence from probing in Web Surveys. In: International Journal of Public Opinion Research. DOI:10.1093/ijpor/eds034

Braun, Michael/Johnson, Timothy P. (2010): An illustrative review of techniques for detecting inequivalences. In: Harkness et al. (Hg): 375-393

Davidov, Eldad/Schmidt, Peter/Billiet, Jaak (Hg.) (2011): Cross-cultural Analysis: Methods and Applications. New York: Routledge

Harkness, Janet A./Braun, Michael/Edwards, Brad/Johnson, Timothy P./Lyberg, Lars/Mohler, Peter P./Pennell, Beth-Ellen/Smith, Tom (Hg.) (2010): Survey Methods in Multinational, Multiregional, and Multicultural Contexts. Hoboken, NJ: Wiley

Harkness, Janet A./van de Vijver, Fons J.R./Mohler, Peter P. (Hg.) (2003): Cross-cultural Survey Methods. Hoboken, NJ: Wiley

Hoffmeyer-Zlotnik, Jürgen H.P./Wolf, Christof (Hg.) (2003): Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables. New York: Kluwer/Plenum

Miller, Kristen/Fitzgerald, Rory/Padilla, José-Luis/Willson, Stephanie/Widdop, Sally/Caspar, Rachel/Dimov, Martin/Grey, Michelle/Nunes, Cátia/Prüfer, Peter/Schöbi, Nicole/Schoua-Glusberg, Alisú (2011): Design and analysis of cognitive interviews for comparative multinational testing. In: Field Methods 23: 379-396

Przeworski, Adam/Teune, Henry (1966): Equivalence in cross-national research. In: Public Opinion Quarterly 30: 551-568

Przeworski, Adam/Teune, Henry (1970): The Logic of Comparative Social Inquiry. New York: Wiley

Rokkan, Stein (Hg.) (1968): Comparative Research across Cultures and Nations. Paris: Mouton

Scheuch, Erwin K. (1968): The cross-cultural use of sample surveys: Problems of comparability. In: Rokkan (Hg.): 176-209

Schuman, Howard (1966): The random probe: A technique for evaluating the validity of closed questions. In: American Sociological Review 31: 218-222

Van Deth, Jan W. (Hg.) (2013[1998]): Comparative Politics: The Problem of Equivalence. ECPR Classics. Colchester: ECPR Press

Van de Vijver, Fons J.R/Leung, Kwok (1997): Methods and Data Analysis for Cross-cultural Research. Thousand Oaks: Sage

Willis, Gordon B. (2005): Cognitive Interviewing: A Tool for Improving Questionnaire Design. Thousand Oaks, CA: Sage